Design and Implementation of a Cross-Media Indexing System for the Reveal-This System

Author 1 Author 2

Abstract

Despite the vast growth of heterogeneous, multimedia and increasingly multi-lingual digital content, there is a lack of integrated technology that facilitates its effective usage. The need is being expressed insistently by end-users, and professionals in content business, particularly, with respect to the re-use and re-purposing of multimedia content. The EU-IST Reveal-This (R-T) project addresses this fundamental need aiming at developing an umbrella technology able to provide retrieval, categorisation, summarisation and translation functionalities that will help users keep up with the explosion of digital content. The system is being developed to cope with digital content that is scattered over different platforms, different languages and different types of media. In order to effectively achieve this vision, the project sets the following two key technical and scientific objectives: 1) benefit from cross-media analysis and indexing techniques and 2) develop representations suitable for cross-media and language technologies that can extend the potential of content push and pull technologies. Research carried out in the project attempts to address all these issues. Inline with the objectives and central to the project, we propose an architectural unit called Cross-Media Indexing Component (CMIC) which constitutes the final layer of the Cross-Media Analysis and Indexing Subsystem. It is employed to cross analyse and leverage the individual potential of each indexing information generated by the analyzers of diverse modalities. In this paper, we focus on CMIC's role and its support for retrieval. The indexing and evaluation approaches are presented together with the status of work in progress.

1 The Reveal-This Project

In the tomorrow's world, people should be spending most of their time enjoying the content, not searching for it. The development of methods and tools for content-based organisation and filtering of large amount of multimedia information is a key issue for effective consumption and enjoyment of digital content. The EU-IST Reveal-This (R-T) project aims at developing a *complete* and *integrated* content programming technology able to capture, index and categorise multimedia content. The project envisions semantic search, summarisation and translation functionalities that will help users cope with vast amount of digital content scattered over different platforms (radio, TV, WWW etc.), different types of media (audio, image, text, video) and different languages.

The R-T system will be used by content providers, to add value to their content, restructure and re-purpose it and offer their subscribers, individual or corporate users a personalised content. For the end users, the system also offers tools for gathering, filtering and categorizing information collected from wide variety of sources in accordance with user preferences.

The project heavily emphasises on promoting cross media analysis and indexing approaches that can extend the possibilites of content push/pull technologies and increasing the quality of content provided. In order to leverage the potential of different modalities, R-T project suggests an architecture which can combine and cross analyse modality specific high-level description of an audio-visual content in a semantic space.

Figure 1 depicts the system architecture of the first R-T system prototype. The Cross-Media Analysis and Indexing subsystem (CAIS) encompasses the state-of-the-art speech, text, image technologies and components for media analysis, indexing and enrichment. Inline with the objectives and central to the project, we propose an architectural unit called Cross-Media Indexing Component (CMIC) which constitutes the final layer of the CAIS. It is employed to cross analyse and leverage the individual potential of each indexing information generated by the analyzers of diverse modalities. In this paper we focus on CMIC's role and its support for retrieval. The indexing and evaluation approaches are presented together with the status of work in progress.

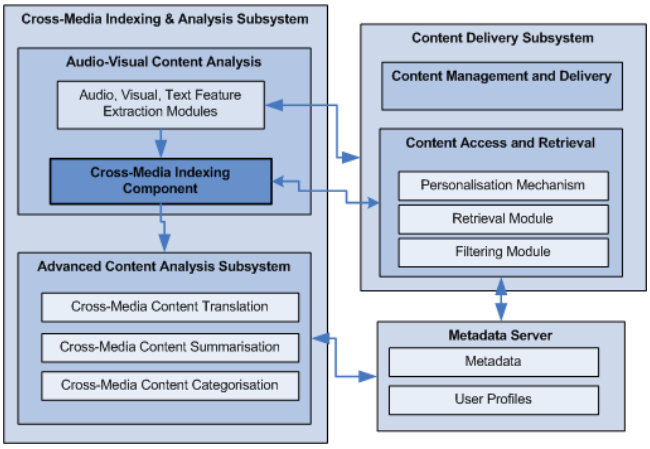

Figure 1. Reveal-This project system architecture

2 Cross-Media Indexing System

The process of cross-media indexing can be defined as building relationships among concepts extracted from different modalities such as image, speech and text which together constitute the multimedia knowledge for a given audio-visual segment. It is the product of recent independent advances made in speech, image and text retrieval research (see for example [18]). A major challenge lies in developing representations suitable for retrieval, categorization and summarisation processes. CMIC attempts to address this challenge by providing necessary software tools to cross analysis and build cross-media indexes. Our research hypothesis is that a system combining these different modal descriptions of the same audio-visual segment can perform better at retrieval and filtering stage. Currently, CMIC utilises the *Multiple Evidence* approach by establishing links among the modality specific descriptions in order to depict topical similarity in the textual space.

The Indexing Process

The process consists of several steps. First the media stream is processed over individual modalities and features are extracted by the speech processing module (speaker turns, gender, transcriptions), the text-processing module (named entities, facts, text-based category) and the image processing module (faces, key-frames, visual features, image categories). Accordingly, segments are identified and regarded as indexable units. The Metadata produced by the above modules are aligned, synchronized and encoded in XML. At this stage, CMIC performs an analysis on the single modality features, enriching the original index with a measure of uncertainty/confidence obtained by comparing the indexing of the different single modalities. As a result, a unified view of the content is produced. This is translated in MPEG-7 [15] for ease of processing and delivering. All modality specific information are mapped to their corresponding MPEG-7 elements; where necessary by extending the standard. Subsequently, this output is handed over to the Cross-Lingual Translation, Cross-Media Summarisation and Content Delivery subsystems that take care of producing a translation in a different language (currently supported languages are English, French and Greek), a personalised and query/profile-biased summary, and delivering the content to the user.

The Indexing Model

The CMIC indexing aims at minimising uncertainty, imprecision and inconsistency across the indexing performed by the single modalities. A robust model is used to reconcile and recognize the relationships among concepts identified in the single modalities. The initial CMIC prototype employs a model based on Dempster-Shafer theory of evidence [16]. The use of this theory has been studied extensively in Information Retrieval [13, 3, 14, 12], but has never been applied in such a context. The framework provides a flexible way of representing and reasoning with imperfect information. Briefly, it combines two or more bodies of evidence defined with in the same *frame of discernment* T into one body of evidence. In our approach, existence of a term t in a document d (segments in our case) provided by a modality m is counted as an *evidence* to support the *topical similarity* hypothesis. Therefore, each analysis module is treated as a *Source of Evidence or Base Probability Assignment* (BPA) defined by a probability density function. BPA contains a set of normalised term frequencies summing to a confidence for each modality m. Instead of treating the cross-media indexing task as a structured document retrieval problem; where evidences propagate to the root node from the leaves, we only combine evidences from modality specific descriptions by applying a modified version of the approach given in [14].

$$m(\emptyset)_d = 0 \text{ and } 1 = \sum_{t \in d} m(\{t\}) \ .$$

where

$$\mathrm{m}_d(\{t\}) = \begin{cases} tf(d,t).\log_N(\frac{N}{n(t)}) & \text{if } t \in d, \\ 0 & \text{otherwise.} \end{cases}$$

$$m_d(T) = 1 - \sum_{t \in d} m_d(\{t\}) \ .$$

$tf(d,t)$ stands for the term t's frequency in document d. $\log_N(\frac{N}{n(t)})$ is a variant of inverse document frequency

where N represents the number of documents in the collection and $n(t)$ represents the number of documents that contain t.

Normalisation can be problematic in some circumstances. However, the preceding formula gracefully overcomes this by $\sum_{t \in d} m_d(\{t\}) \leq 1$ and adheres to the theory. $m(T)$ is equal to the unassigned BPA to any preposition set or the hypothetical uncertainty. Instead of treating the cross-media indexing task as a structured document retrieval problem [9]; where evidences propagate to the root node from the leaves, we only combine evidences from modality specific descriptions by applying a modified version of the following combination rule [14]:

$$m(\{t\}) = m_1 \otimes m_2(\{t\}) = \frac{1}{K}(m_1(T).m_2(T))$$

where

$$\mathrm{K} = \left(\sum_t m_1(\{t\}).m_2(\{t\}) \right) + m_1(T).m_2(T) \ .$$

such that $m_1(\{t\}) > 0 \ and \ m_2(\{t\}) > 0$ conditions are satisfied.

3 Related Work

Cross-media indexing and retrieval is a product the recent advances in each speech, image and text retrieval research. In multimedia retrieval research, it has been long studied that, a single modality can be analyzed and indexed in a certain way which would effectively increase retrieval system's performance [1].

The research in speech retrieval [2, 6, 5] has shown that, even under high levels of word recognition error rates, it is possible to get reasonable retrieval performance using classical information retrieval techniques. In [17], authors focus on expanding documents to alleviate the effect of transcription mistakes on speech retrieval. Discussion of the effects of out of vocabulary items in spoken document retrieval is given in [20]. Experiments on TREC-8 (1998) audio collection using various retrieval setups suggests that it is possible to demonstrate moderate retrieval performance when advanced IR techniques (a combination of Query and Document expansion method) are used to compensate for recognition errors caused by out of vocabulary words. Experiments include query expansion and document expansion against a baseline retrieval system that uses Okapi variant.

More recently, research has been moving forward to cross modal analysis of media information. Recent advances in the image retrieval report that text (in the form of annotations or speech transcripts) and image (histogram, texture, colour frequencies etc.) can be combined successfully [4, 7, 19]. In [4], authors suggest an approach for predicting words that are associated to whole images and corresponding to particular image regions. The process is regarded as a problem of translation, such as translation of image regions to words. Using multi-modal and correspondence extensions to Hofmann's [10] hierarchical clustering/aspect model (which is a model adapted from statistical machine translation and a multi-modal extension to latent dirichlet allocation) is proposed. In [8], Duygulu extends the previous work and studies videos rather than images. This is a difficult problem since, text is not associated with a single frame. Authors report reasonable success over TREC2001 collection. In another study [1], semantic labeling is formulated as a machine learning problem. Concept representations are modeled using Gaussian mixture models(GMM), Hidden Markov Models(HMM), and support vector machines(SVM) and were evaluated against TrecVID(2001) corpus. Study reports that fusion scheme achieves more than %10 relative improvement over the best single modal concept detectors.

4 Evaluation

In order to validate the first prototype, we intent to evaluate the performance of different cross-media indexing models. This will enable us to find the best model for the combination of evidence. In this context we are currently pursuing two different strategies, both involving the construction of a test collection using real users. The first one is *known item search*. The users are given a collection of multimedia segments and are asked to write a query that will find a very specific segment. This simple process makes it easy to build quickly a test collection to be used in the preliminary phases of evaluation. At the same time we are building a proper *test collection*. The users are given a test video sequence and then asked to detect-recognize faces, transcribe images and describe other aspects of the context. Their descriptions are regarded as the ground truth. Later on, search and filtering tasks are introduced. At this stage, a user describes information needed to accomplish a specific work related task and exhaustively go through the collection to identify the segments that are relevant to the specific need. Here the challenge is preparing and building the collection with relevance judgments easily and quickly. The availability of this test collection will enable to test not only CMIC, but also the reliability of the single modality indexing modules. CMIC will make use of the results of the evaluation of the single modality indexing module for weighting the reliability of the evidence provided by each of them.

The collection we are using for both evaluation strategies is three-hour multimedia collection. The collection covers politics, travel and news domains in English, French and

[1] An up-to-date survey of the research fields can be found in [18]

Greek languages.

It should be noted that the final evaluation of the R-T system will be carried out using a user and task oriented evaluation involving home user and TV broadcast professionals.

5 Conclusions and Future Work

We are on the way to show that the *Multiple Evidence* approach can be employed to robustly reconcile a unique and complete description of the topical content which will help users find relevant information compared to single modal and separate indexing strategies. We are exploring other combination rules such as Yager's modified Dempster's Rule [21] and Inagaki's unified combination rule [11]. In addition, we also intend to experiment with more advanced approaches based on Bayesian Networks, Kernel Canonical Analysis, Gaussian Mixture models.

References

[1] W. H. Adams, G. Iyengar, C.-Y. Lin, M. R. Naphade, C. Neti, H. J. Nock, and J. R. Smith. Semantic indexing of multimedia content using visual, audio, and text cues. *Journal on Applied Signal Processing*, 2003(2):170185, 2003.

[2] J. Allan. Perspectives on information retrieval and speech. In *SIGIR Workshop: Information Retrieval Techniques for Speech Applications*, pages 1–10. Springer, 2001.

[3] Y. A. Aslandogan and C. T. Yu. Multiple evidence combination in image retrieval: Diogenes searches for people on the web. In *Proceedings of the ACM SIGIR*, Athens, Greece, 2000. ACM Press.

[4] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.

[5] F. Crestani. Spoken query processing for interactive information retrieval. *Data and Knowledge Engineering*, 41(1):105–124, 2002.

[6] F. Crestani. Combination of similarity measures for effective spoken document retrieval. *Journal of Information Science*, 29(2):87–96, 2003a.

[7] A. P. de Vries, T. Westerveld, and T. Ianeva. Combining multiple representations on the trecvid search task. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004)*, 2004.

[8] P. Duygulu, D. Ng, N. Papernick, and H. Wactlar. Linking visual and textual data on video. In *Workshop on Multimedia Contents in Digital Libraries*, Crete, Greece, 2003.

[9] A. Graves and M. Lalmas. Video retrieval using an mpeg-7 based inference network. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, Tampere, Finland, 2002. ACM Press.

[10] T. Hofmann. Learning and representing topic. a hierarchical mixture model for word occurrence in document databases. In *Workshop on learning from text and the web*, CMU, 1998.

[11] T. Inagaki. Interdependence between safety-control policy and multiple-sensor schemes via dempster-shafer theory. *IEEE Transactions on Reliability*, 40(2), 1991.

[12] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, Toronto, Canada, 2003. ACM Press.

[13] J. M. Jose and D. J. Harper. A retrieval mechanism for semi-structured photographic collections. In *Proceedings of the DEXA*. Springer-Verlag, 1997.

[14] M. Lalmas and E. Moutogianni. A dempster-shafer indexing for the focussed retrieval of a hierarchically structured document space: Implementation and experiments on a web museum collection. In *Proceedings of RIAO*, Paris, France, 2000.

[15] J. M. Martnez. Mpeg-7: Overview of mpeg-7 description tools. *IEEE Multimedia*, 9(3):83–93, 2002.

[16] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[17] A. Singhal and F. Pereira. Document expansion for speech retrieval. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, Berkeley, California, United States, 1999. ACM Press.

[18] C. Snoek and M. Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications*, 25(1):5–35, 2005.

[19] T. Westerveld. *Using generative probabilistic models for multimedia retrieval*. PhD thesis, 2004.

[20] P. C. Woodland, S. E. Johnson, P. Jourlin, and K. S. Jones. Effects of out of vocabulary words in spoken document retrieval (poster session). In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, Athens, Greece, 2000. ACM Press.

[21] R. Yager. On the dempster-shafer framework and new combination rules. *Information Sciences*, 41, 1987.